

# historical text mining historical text mining, and historical text mining: challenges and opportunities

Dr. Robert Sanderson  
Dept. of Computer Science  
University of Liverpool

azaroth@liv.ac.uk  
<http://www.csc.liv.ac.uk/~azaroth/>

- ▶ Text Mining?
- ▶ 'History' of Text Mining
- ▶ Mining Texts about History
- ▶ Mining Historical Texts
- ▶ Opportunities for the Future

- ▶ Text Mining: No canonical definition
- ▶ Commonly used definition based on Data Mining:

“The non-trivial extraction of implicit, previously unknown, and potentially useful information from data.”



“The non-trivial extraction of previously unknown, interesting facts from a collection of texts.”

Isn't this just Data Mining with Text?

- ▶ Typical Data Mining Functions:
  - ▶ Classification
  - ▶ Association Rule Mining
  - ▶ Clustering

Useful when applied to texts, but doesn't fulfill the definition as they don't discover “facts”.

Isn't this just Information Retrieval?

- ▶ Information Retrieval:
  - ▶ Locates most relevant documents to query
  - ▶ Returns documents
  - ▶ Clusters documents (etc)

Still doesn't fulfill the definition as it doesn't discover “facts”, it works only at the document level.

Need to understand the meaning of the text:

- ▶ Part of Speech tagging
- ▶ Deep Parsing (clauses)
- ▶ Named Entity Recognition
- ▶ Information Extraction
- ▶ Infer information from correlations

Result: New Knowledge

Plus a lot more:

- ▶ Improved document classification
- ▶ Automatic semantic annotation of documents
- ▶ Improved access -- search by semantics and concepts
- ▶ Improved clustering of documents by concept
- ▶ Summarization
- ▶ Visualization techniques

- ▶ Information Processing Framework
- ▶ Standards based: XML, SRU, Unicode, etc.
- ▶ Scalable: Single machine to Grid (PVM, MPI, SRB)
- ▶ Extensible: Python + C, Object Oriented with stable API
  
- ▶ Tools from NaCTeM partners integrated
- ▶ Medline: 4,350 records/second using 60 processes at SDSC
- ▶ Bibliographic: 440 seconds for 16 million records (1 field)



**Parmenides Common Annotations Manager**

File Actions Options View Help

Business Group BALTIMORE, Jan. 5  
 ViPS, a leading provider of business intelligence for the healthcare and life science industries, today announced the appointment of Karen Borda to vice president and general manager. In her new role at ViPS BioMedical Services, Borda will manage all business development and operations of current products and services for the group. She will also spearhead the expansion of value-added products as well as data and consulting services.  
 "Karen has excellent knowledge of the clinical trial process and strong operational and technological experience," states Jenny Morgan, president and chief executive officer at ViPS. "Her solid achievements in operations, business processes and software systems paired with her strong management background give her extensive insight into the opportunities and challenges ViPS BioMedical Services faces."  
 Borda's career in life sciences spans 18 years, the last 10 of which Borda has served in senior executive positions. Before joining ViPS, Borda was a partner at Quadragen, Inc., a consulting firm specializing in the pharmaceutical industry. Borda was also a founding partner of CB Technologies, where she was instrumental in the initial success of the company.  
 "Given the market demand for proven business intelligence systems and ViPS' long history of effective delivery, I'm excited about the opportunity to help drive the strategic direction of ViPS BioMedical Services," states Borda. "We'll be aggressively growing our leadership position in life science solutions, dependable ROI and quick implementation."  
 In June 2003, ViPS acquired the former CB Technologies to be its strategic platform for building pharmaceutical-based businesses. Today ViPS BioMedical Services remains the premier provider of technological tools and services to the life science industries - helping pharmaceutical, biotechnology, medical device and contract research organizations adopt more efficient processes in clinical trials. The company's state-of-the-art electronic data capture (EDC) software, exceptional service and unparalleled technology integration have been supporting life science

**Lexical Annotations**

- [-] Namexes
  - URL
  - company
  - position
  - person
  - product
  - service
  - event\_head
  - event\_phrase
  - not\_event\_noun
- [-] Timexes
- [-] Entities
  - position
  - product
  - event\_phrase
  - company
  - url
  - person
    - Borda
    - Morgan
  - product1
  - service
  - not\_event\_noun
  - event\_head
  - product2
- [-] Relations
- [-] Events
  - joining\_event
  - personnel\_event
  - acquisition\_event

**Entity**

ID	Type	Mnem	References
pe17	person	Borda	pn42 pn47 pn48 pn49 pn45 pn44 pn46

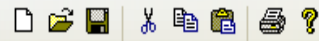
**Entities of type [person]**

ID	Type	Mnem	References
1	person	Borda	pn42 pn47 pn48 pn49 pn45 pn44 pn46
2	person	Morgan	pn43

D:\CVSFiles\resources\CaseStudies\Biovista\BV\Axsl\Corpus\BV\Apersonnel\BV\Apersonnel001.xml

TerMine [GUI]
\_ □ ×

File Edit View Help



Source documents

**Beta-arrestin binding to the beta2-adrenergic receptor requires both receptor phosphorylation and receptor activation.**

Krasel C, Bunemann M, Lorenz K, Lohse MJ.  
Institute for Pharmacology and Toxicology, Versbacher Strasse 9, D-97078 Wurzburg, Germany.

Homologous desensitization of beta2-adrenergic receptors has been shown to be mediated by phosphorylation of the agonist-stimulated receptor by G-protein-coupled receptor kinase 2 (GRK2) followed by binding of beta-arrestins to the phosphorylated receptor. Binding of beta-arrestin to the receptor is a prerequisite for subsequent receptor desensitization, internalization via clathrin-coated pits, and the initiation of alternative signaling pathways. In this study we have investigated the interactions between receptors and beta-arrestin2 in living cells using fluorescence resonance energy transfer. We show that (a) the initial kinetics of beta-arrestin2 binding to the receptor is limited by the kinetics of GRK2-mediated receptor phosphorylation; (b) repeated stimulation leads to the accumulation of GRK2-phosphorylated receptor, which can bind beta-arrestin2 very rapidly; and (c) the interaction of beta-arrestin2 with the receptor depends on the activation of the receptor by agonist because agonist withdrawal leads to swift dissociation of the receptor-beta-arrestin2 complex. This fast agonist-controlled association and dissociation of beta-arrestins from prephosphorylated receptors should permit rapid control of receptor sensitivity in repeatedly stimulated cells such as neurons.

**beta2-adrenergic receptor gene single-nucleotide polymorphisms are associated with rheumatoid arthritis in northern Sweden.**

Xu B, Arlehang L, Rantapaa-Dahlquist SB, Lefvert AK.  
Department of Immunology, American Red Cross Biomedical Research and Development, MD 20855, USA. xubiy@usa.redcross.org

The beta2-adrenergic receptor (beta2-AR) belongs to the group of G-protein-coupled receptors and is present on skeletal and cardiac muscle cells and on lymphocytes. The gene encoding beta2-AR (ADRB2) displays a moderate degree of heterogeneity in the human population and the distributions of single-nucleotide polymorphisms (SNPs) at amino acid positions 16, 27, and 164 are changed in asthma, obesity, and hypertension and in the autoimmune disease myasthenia gravis. An involvement of the beta2-AR has also been suggested in human rheumatoid arthritis (RA) and its animal model. We describe here an increased prevalence of the alleles Arg16 and Gln27 and a lower prevalence of homozygosis for Gly16 and Glu27 in patients with RA. Patients having the genotype combination GlyGly16-GlnGlu27 had higher levels of rheumatoid factor (RF) and a more active disease than other patients. Patients having the genotype Arg16-Gln27+ had higher levels of RF when compared to those having Arg16+Gln27+, and patients who were carriers of Gln27 had a more active disease than non-carriers of Gln27. Our results show an association of beta2-AR SNPs with RA in a population from the northern part of Sweden. Our study also confirms the strong linkage disequilibrium of genotypes at amino acid

Result 1 - 50 of about 1131 terms

Rank	Term	Score
1	beta2-adrenergic receptor	65.7778
2	blood pressure	16.8
3	beta2-adrenergic receptor gene	14.8496
4	single nucleotide polymorphisms	9.50977
5	adrenergic receptor	9.14286
6	Gly16 allele	8
7	A549 cells	8
8	body mass index	7.92481
9	cystic fibrosis patients	7.92481
10	protein kinase	7.625
11	cystic fibrosis	7.33333
12	confidence interval	7
13	metabolic syndrome	7
14	allelic frequency	7
15	bioluminescence resonance energy transfer	6.8

Ready

CytoSailing - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www-tsuji.is.s.u-tokyo.ac.jp/CytoSailing/

Go

My Folder 1

- >> Folder
- brc-1 +memo
- BIN2 +memo

Gene Dictionary

Search for genes or gene product names

>> Organism >> Field

CAS2	breast carcinoma amplified sequence 2	spliceosome associated protein, amplified in breast cancer	
	breast carcinoma amplified sequence 3	metastasis associated antigen of breast cancer	
	Breast cancer-related regulator of TP53		
	bridging integrator 2	breast cancer	

Interaction Viewer

>> Change settings | < >

Drag interaction objects into *Content Viewer* to see their evidence sentences.

Entity	Object	Match
BIN2	ITGA2	2 / 7  BIN2  ITGA2
<input type="button" value="next"/>	BIN3	0 / 5  BIN2  BIN3
<input type="button" value="Erase"/>	TFF1	0 / 3  BIN2  TFF1
	BRCA1	0 / 2  BIN2  BRCA1
	GYS1	0 / 2  BIN2  GYS1
	CESK1	0 / 2  BIN2  CESK1
	BAK1	0 / 1  BIN2  BAK1
	BARD1	0 / 1  BIN2  BARD1
	GSK3B	0 / 1  BIN2  GSK3B
	PRV1	0 / 1  BIN2  PRV1

Content Viewer 1

>> Folder | < >


Symbol	brc-1
Name	BReast and ovarian Cancer susceptibility protein homolog (62.3 kD) (brc-1)
Organism	Caenorhabditis elegans
Link	
Synonym	C36A4.7 / C36A4.8 / CELK07078
Product	BReast and ovarian Cancer susceptibility protein homolog (brc-1)

associated protein AP1	
adder	
cancer-associated tein	
adder cancer	
ociated protein	
adder cancer related tein	
adder cancer related tein (10kD)	
ast and ovarian cancer susceptibility tein 1	

Done 6.301s Adblock

Hilragi - Semantic retrieval engine for MEDLINE - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.tsujii.is.s.u-tokyo.ac.jp/hilragi/ Go 

Semantic Search **Keyword Search** GCL Search Custom Search User Profile

subject verb object

kill  cancer Search! Clear Help


»Keyword list  
»Advanced search

---

Results 1-50 for **kill cancer** »Show next »Show query 15.47 seconds (3.81% finished)

Query: [sentence sentence\_id="\$sentence"] >> ([word arg1="\$subject" arg2="\$object" id="\$verb" base="kill"] & ([phrase cat="np" id="\$object"] > ([word id="\$object\_match1"] >> cancer)))) «Hide

- [PMID: 11772236](#) »XML  
More importantly , DNA-damaging anticancer agents , such as adriamycin , kill cancer cells , at least in part , by upregulating FasL .
- [PMID: 11791373](#) »XML  
They assert that the aim of low-dose chemotherapy is to prolong time to progression ( TTP ) , not to kill the cancer cells , so it may reasonably be called " tumor dormancy therapy " .
- [PMID: 11798956](#) »XML  
CONCLUSION : Peritoneal lavage , helpful in killing the exfoliated cancer cells in peritoneal cavity of patients with gastric cancer at II and III A stages , should be conducted in the treatment of gastric cancer by radical gastrectomy before closing the abdomen .
- [PMID: 11850720](#) »XML  
We recently showed that the human telomerase reverse transcriptase ( hTERT ) promoter induces tumor-specific Bax gene expression and selectively kills various human cancer cells both in vitro and in xenograft tumors .
- [PMID: 11850832](#) »XML  
TRAIL is a pro-apoptotic cytokine believed to selectively kill cancer cells without harming normal ones .
- [PMID: 11857034](#) »XML  
Full-length TRAIL delivered by an adenoviral vector ( AdTRAIL-IRES-GFP ) killed prostate cancer cell lines and PrEC without requisite doxorubicin cotreatment .
- [PMID: 11877684](#) »XML  
Targeted toxins represent novel cancer therapeutics designed to selectively target and kill cancer cells .

Done 2.224s Adblock 

- ▶ Extract events from the text along with information about the participants
- ▶ Can be modeled as relationships between named entities
- ▶ Extracting events allows discovery of hidden temporal correlations  
eg: Google refuses to announce plans. Google's stock falls.
- ▶ Improves understanding of the semantics, improving the functions based around those semantics

- ▶ Focus on current events and bioinformatics.
- ▶ Why?
- ▶ Easily available large scale data sets
- ▶ Easily available remote services and tools
- ▶ Easily available ontologies and thesauri
- ▶ Easily available \$\$\$ £££ ¥¥¥ €€€

- ▶ Today's news is tomorrow's history
- ▶ People, Events, Places, Dates
- ▶ a.k.a: Who, What, Where, When
- ▶ Want to find out correlations ... a.k.a Why

## Challenges:

- ▶ Entity recognition harder due to ...
- ▶ Lack of ontologies due to ...
- ▶ Very broad scope (centuries compared to days)

- ▶ Fewer easily available datasets

Need ArXiv, CiteSeer or Medline for History

OTA? Project Gutenberg? (Brown, Cobuild, BNC?)



- ▶ Language

  - Most tools focus on English (Bio, CompSci)

  - History in all languages, dead and alive

  - TreeTagger, Chasen, Juman...

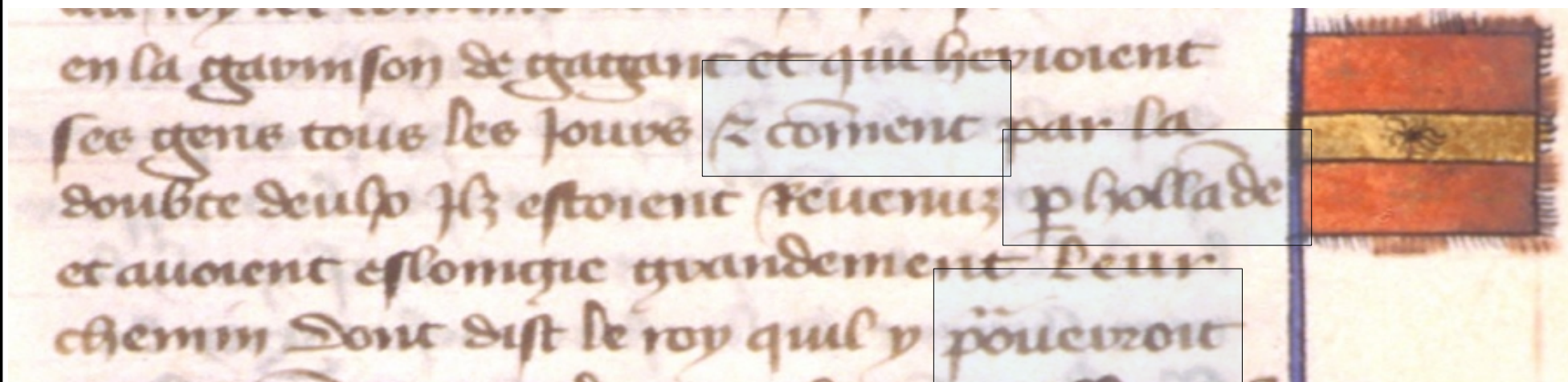
- ▶ Services, Tools:

  - No remote services?

- ▶ Need OSS or remotely available taggers, ontologies etc!

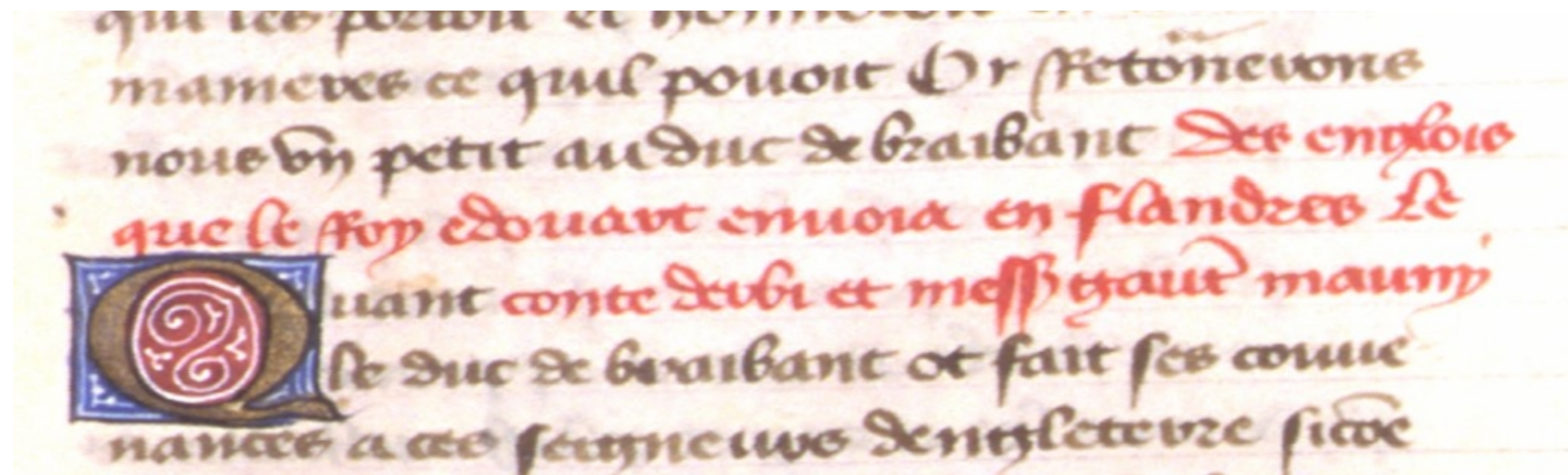
- ▶ Datasets? Tools? Services? Anything?!
- ▶ Perseus Project, Cultural Heritage Language Technology
- ▶ Much harder across the board
- ▶ Additional issues need additional experts from different areas
- ▶ Difficult to get full transcriptions of the texts to work with
- ▶ Histoire versus Chronique – hard to extract events
- ▶ History written by ... the most frequently copied manuscripts

## ► Abbreviations



Needed: Smart abbreviation dictionary, standardised way to record abbreviations, re-rendering mechanisms?

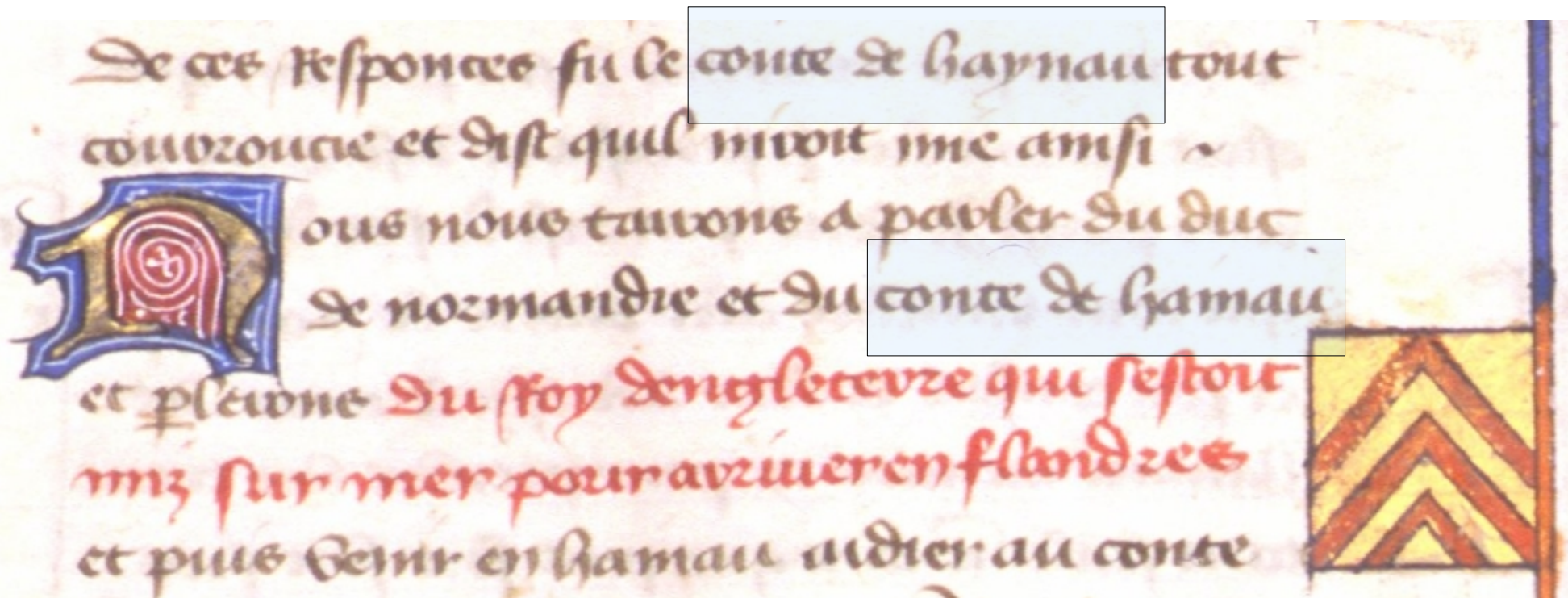
## ► Broken Text Flow



Needed: Language models + image analysis

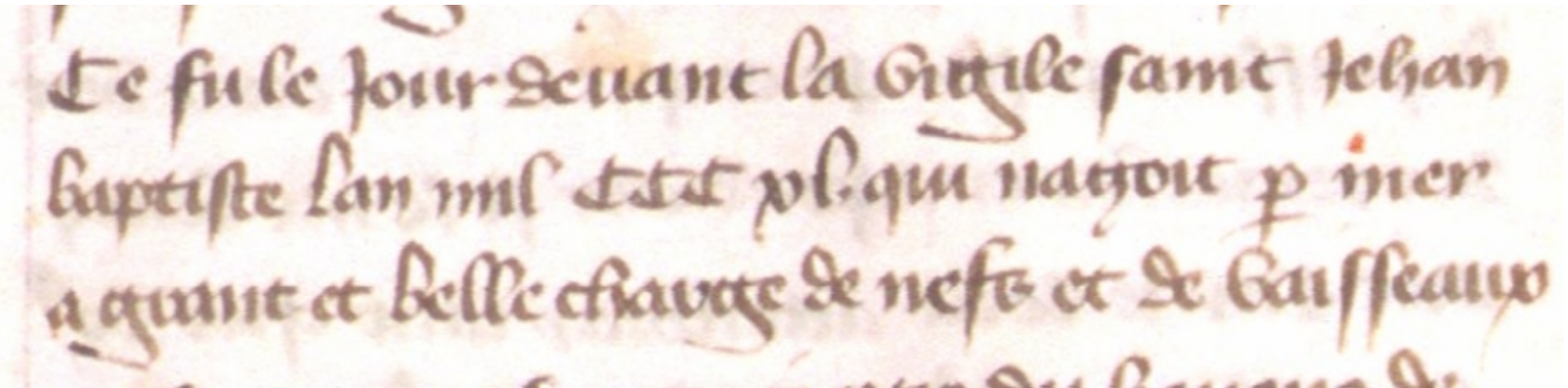
Note: 'Q' as illuminated character within Quant

- ▶ Little concept of 'correct' spelling



Needed: Language models, smart gazeteers, smart dictionaries

► Date recognition



Needed: Smart chronological tools

- ▶ Increase availability of data sets, tools, services
- ▶ Increase the availability and production of ontologies, thesauri, taxonomies, gazeteers, chronologies, dictionaries...
- ▶ ... and link them in with language models.
- ▶ Identifiers! Identifiers! Identifiers!
- ▶ Text Mining as an aid to transcription
- ▶ Most importantly: Collaboration!

# Thank You

---

▶ Questions?